Psychological Medicine (2012), **42**, 647–656. © Cambridge University Press 2011
doi:10.1017/S0033291711001292

ORIGINAL ARTICLE

Validating two survey methods for identifying cases of autism spectrum disorder among adults in the community

T. S. Brugha[1]*, S. McManus[2], J. Smith[1], F. J. Scott[3], H. Meltzer[1], S. Purdon[2], T. Berney[4], D. Tantam[5], J. Robinson[6], J. Radley[7] and J. Bankart[1]

[1] *Department of Health Sciences, University of Leicester, UK*
[2] *National Centre for Social Research (NatCen), London, UK*
[3] *Autism Research Centre, Department of Psychiatry, University of Cambridge, UK*
[4] *Institute of Health and Society, Newcastle University, UK*
[5] *School of Health and Related Research, University of Sheffield, UK*
[6] *Cambridgeshire and Peterborough NHS Foundation Trust (CLASS Clinic), UK*
[7] *St Andrews Healthcare, Northampton, UK*

Background. There are no tested methods for conducting epidemiological studies of autism spectrum disorders (ASDs) in adult general population samples. We tested the validity of the Autism Diagnostic Observation Schedule module-4 (ADOS-4) and the 20-item Autism-Spectrum Quotient (AQ-20).

Method. Randomly sampled adults aged ⩾16 years were interviewed throughout England in a general population multi-phase survey. The AQ-20 was self-completed by 7353 adults in phase 1. A random subset completed phase 2, ADOS-4 assessments ($n=618$); the probability of selection increased with AQ-20 score. In phase 3, informant-based Diagnostic Interview Schedule for Social and Communication Disorders (DISCO) and Autism Diagnostic Interview – Revised (ADI-R) developmental assessments were completed ($n=56$). Phase 1 and 2 data were presented as vignettes to six experienced clinicians (working in pairs). The probability of respondents having an ASD was compared across the three survey phases.

Results. There was moderate agreement between clinical consensus diagnoses and ADOS-4. A range of ADOS-4 caseness thresholds was identified by clinicians: 5+ to 13+ with greatest area under the curve (AUC) at 5+ (0.88). Modelling of the presence of ASD using 56 DISCO assessments suggested an ADOS-4 threshold in the range of 10+ to 13+ with the highest AUC at ADOS 10+ to 11+ (0.93–0.94). At ADOS 10+, the sensitivity was 1 [95% confidence interval (CI) 0.59–1.0] and the specificity 0.86 (95% CI 0.72–0.94). The AQ-20 was only a weak predictor of ADOS-4 cases.

Conclusions. Clinically recommended ADOS-4 thresholds are also recommended for community cases: 7+ for subthreshold and 10+ for definite cases. Further work on adult population screening methods is needed.

Received 8 February 2011; Revised 9 June 2011; Accepted 21 June 2011; First published online 29 July 2011

Key words: Adult, autistic disorder, epidemiologic methods, instrumentation, validity and reliability.

Background

Autism spectrum disorders (ASDs) are developmental disorders characterized by impairment of reciprocal social interaction, social communication and social imagination, often in the presence of restricted repetitive behaviours (Wing, 1997), with negative impacts on learning and independence in adulthood (Howlin *et al*. 2004), affecting individuals with abilities ranging from the profoundly learning disabled to the intellectually superior, of whom some may also suffer from other co-morbid psychiatric disorders. The cost to society, individuals and families of ASDs in the UK is estimated to be approximately £90 000/year for each adult with ASD (Knapp *et al*. 2007). Until now, epidemiological surveys of ASD have only been conducted using samples of children (Fombonne 2003, 2005; Baird *et al*. 2006; Baron-Cohen *et al*. 2009). Understanding the epidemiology and causes of adult-onset co-morbid severe mental disorders such as schizophrenia and bipolar disorder (Cichon *et al*. 2009) will also require accurate phenotyping of autism in adult community samples.

The concept of autism was first written about in accessible form in the mid-twentieth century (Kanner, 1943; Asperger, 1991) and is still evolving (Frith,

* Address for correspondence: Professor T. S. Brugha, Department of Health Sciences, University of Leicester, Leicester General Hospital, Leicester LE5 4PW, UK.
(Email: tsb@le.ac.uk)

1991). Experts have achieved a consensus on what constitutes the category of Autism Spectrum Disorder (ASD), also known as Pervasive Developmental Disorder (PDD) (WHO, 1993; APA, 1994). Two issues hamper diagnosis in adulthood. First, information on childhood development and behaviour (required by both DSM-IV and ICD-10) is frequently unavailable. Second, clinical experience shows that adults with ASD are often unable to describe their own social difficulties and behaviour.

Methods for conducting large epidemiological surveys of mental disorder among adults in the general population rely on direct interviews with survey respondents by lay interviewers (Brugha & Meltzer 2008). Questions address subjective emotional, cognitive and physical states and rarely include information on observed behaviour, which would require an informant such as a partner or carer. Such surveys have often incorporated clinical diagnostic assessments by face-to-face or telephone interview (Brugha *et al*. 1999; Haro *et al*. 2006) after a fully structured lay interview and/or completion of a self-report questionnaire. Comparison of these phase 1 and 2 assessments can determine the most clinically appropriate diagnostic threshold on the survey phase 1 self-report measure. Thresholds used by clinicians making diagnoses in practice also have a role in instrument standardization.

In surveys of adults, the ideal scenario would involve standardized assessments of directly observed (current) behaviour of adults using, for example, the Autism Diagnostic Observation Schedule module 4 (ADOS-4; Lord *et al*. 2002). Information from parents on early development and on current day-to-day functioning over a representative period of adult life in the community could also be sought. Potentially suitable standardized instruments for this purpose are the Autism Diagnostic Interview – Revised (ADI-R; Lord *et al*. 1994) and the Diagnostic Interview Schedule for Social and Communication Disorders (DISCO; Wing *et al*. 2002), together, perhaps, with a sample of judgements by clinicians. However, such informant-dependent and intensive approaches are not a viable proposition in large-scale adult general population surveys, particularly for the older age range. Self-report questionnaires such as the Autism-Spectrum Quotient (AQ; Baron-Cohen *et al*. 2001) or a structured diagnostic interview could be used but must first be tested and calibrated using more intensive assessment methods. Elsewhere we describe the derivation and testing of a 20-item version of the AQ, the AQ-20, for use in surveys (T. S. Brugha *et al*., unpublished observations).

The aim of the current study was to validate a clinical diagnostic measure for estimating the prevalence and describing the epidemiology of ASD in adults in the community, the ADOS-4, using standardized developmental assessments and ratings by clinicians, and to test the sensitivity and specificity of a survey self-report screening measure, the AQ-20.

Method

Two methodology studies were carried out as an extension to the third Adult Psychiatric Morbidity Survey (APMS) by the National Centre for Social Research (NatCen) in collaboration with the University of Leicester (Brugha *et al*. 2009, 2011; McManus *et al*. 2009). In Study 1 we tested and calibrated the ADOS-4 as an observer-based diagnostic measure. In Study 2 we tested the sensitivity and specificity of the AQ-20 used as a short self-report questionnaire measure of ASD in adults in the community (Brugha *et al*., unpublished observations). A three-phase general population survey design was used (Fig. 1).

Survey design

Phase 1 data (Fig. 1) were obtained from a random probability sample of the general population of England, as described previously (Brugha *et al*. 2009, 2011; McManus *et al*. 2009). Sampling of primary sampling units (PSUs) was followed by the sampling of addresses within the selected PSUs. Interviewers visited the 14 532 addresses to identify private households with at least one person aged $\geqslant$16 years. A total of 13 114 addresses (90.1%) were found to contain private households. One person per household was randomly selected to take part in the survey. If the selected respondent was incapable of undertaking the interview, for reasons of mental or physical incapacity, a 'proxy' interview was permitted with another person who knew the selected respondent well but no information relevant to ASD was collected.

For each phase 1 participating respondent (n=7353), the probability of selection for a phase 2 assessment (Fig. 1) was calculated as the maximum value of four disorder-specific probabilities: psychosis probability, ASD probability, borderline personality disorder probability and antisocial personality disorder probability. Thus the probability of selection increased with AQ-20 score (McManus *et al*. 2009). The phase 2 interviews carried out were ADOS-4 (Lord *et al*. 2002), the survey format of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN; Brugha & Nienhuis, 1998; Wing *et al*. 1990) and the SCID-II semi-structured interview (Williams *et al*. 1992).

For the present study, a phase 3 sample was drawn at random from eligible respondents who had

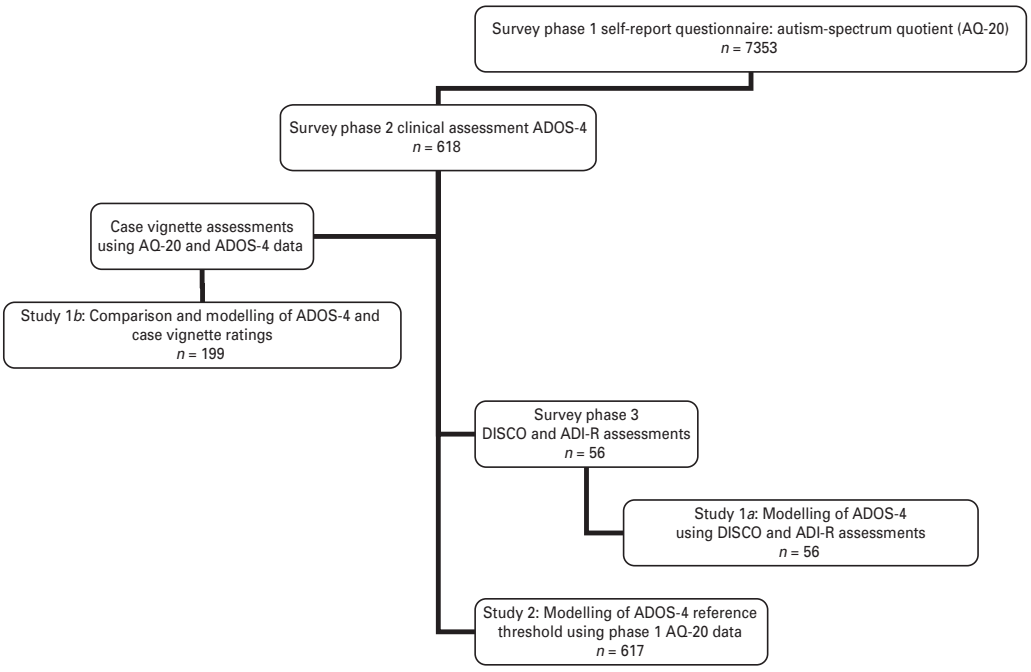

Fig. 1. Outline of study design.

completed both the first and second phases of the APMS (Fig. 1). Respondents must have given consent and an informant such as a partner, carer or parent had to be available and willing to take part. Thirty respondents with a high ADOS-4 score were selected (ADOS-4 Communication >1; Reciprocal Social Interaction >3; total combined score >6) and a randomly selected group of 30 who were negative on the ADOS-4 (controls; ADOS-4 <7). To assess agreement between the phase 2 ADOS-4 and the phase 3 DISCO and ADI-R, 60 people who complete these two assessments would provide sufficient precision to find a κ of 0.6 with a 95% confidence interval (CI) extending 0.22 in either direction, assuming a two-tailed hypothesis, assuming an expected true proportion of successes of 70% (nQuery v. 2.0; Elashoff, 1997).

Instruments

The 20-item version of the AQ was used throughout phase 1 (McManus *et al*. 2009; Brugha *et al*., unpublished observations). In collaboration with the developers of the original 50-item AQ (Baron-Cohen *et al*. 2001), the 20-item version, AQ-20, was derived using regression model methods in a sample of adults referred to a specialist ASD clinic and student volunteers (Brugha *et al*., unpublished observations). The AQ-20 gave an area under the curve (AUC) of 0.983, with a sensitivity of 86% and a specificity of 99%. The AQ-20 was as good a predictor as the AQ-50 in the sample from which it was derived.

The ADOS-4 provides a direct face-to-face assessment of current respondent behaviour consistent with a diagnosis of an autistic disorder (Lord *et al*. 2002). It consists of various tests comprising set situations (termed 'presses') that evaluate communication, reciprocal social interaction, creativity, imagination and stereotyped behaviour and restricted interests. Respondents are also asked about their knowledge and understanding of social relationships, emotions and daily living responsibilities. Algorithms for ASD and for autism are incorporated in the ADOS-4 (Lord *et al*. 2002). Selected ADOS-4 ratings that correspond to DSM-IV and ICD-10 criteria for PDD are summed to a total score for Communication and Reciprocal Social Interaction to which two thresholds are applied for non-specific PDD ($\geqslant 7$ on the ADOS-4 combined total score: we have termed 'ADOS 7+') and for Autism ($\geqslant 10$ on the ADOS-4 combined total score: termed 'ADOS 10+').

A training programme for interviewers on ADOS-4 was developed according to ADOS principals (Lord *et al*. 2002) using typical examples of abnormal behaviours likely to be encountered in fieldwork. Volunteers taking part usually had a clinician-determined diagnosis of ASD. There was little difficulty in discussing and agreeing on ADOS-4 ratings when considered in the context of independent adults living unsupported in the community. For example, when rating 'Quality of Social Overtures', which is a rating of the quality of the respondent's attempts to initiate social interaction with the examiner, it is stated

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

(Lord *et al*. 2002) that 'special attention should be given to the form of the overture and its *appropriateness to the social context*' [italics added]. Fieldwork interviewing did not commence until the team of four interviewers was achieving at least 90% agreement on ratings of jointly observed ADOS-4 examinations using the thresholds 0, 1, 2 or more.

The ADOS is only one component of a diagnostic assessment that should include information from an informant (Lord *et al*. 2002). Information on adult functioning was provided by a parent, sibling, partner or current carer. The DISCO was used in phase 3 (Fig. 1) to generate classifications of a range of possible developmental disorders [e.g. ASD, tourettes, attention deficit hyperactivity disorder (ADHD), general ability level] based on detailed information on early development (if available) and on current and recent behaviour (Wing *et al*. 2002). The ADI-R (Lord *et al*. 1994) was also coded at the same time by the same interviewer for comparison purposes. The ADI-R, like the DISCO, is considered to be a reference or 'gold standard' research assessment of early development and current behaviour, but on the autism spectrum only.

Fieldwork procedures, training and quality control

The ADOS protocol recommends audiovisual recording to facilitate rating (Lord *et al*. 2002). Although this was included in the ethical approval for the APMS, it proved too demanding to implement for respondents in their own homes (other than for some pilot assessments). Quality control measures were built into the survey process, both at data collection and to check on the quality of phase 1 and phase 2 interviewer performance. The phase 2 interview is less structured and requires clinical skills and judgement. The fieldwork of the research psychologists was supervised by a senior research psychologist and a psychiatrist who had conducted earlier surveys in the GB Psychiatric Morbidity programme (Jenkins *et al*. 2009). At the midpoint of phase 2 fieldwork, all interviewers met again with the ADOS trainer and a second equally experienced ADOS trainer who had not been involved in their training to conduct an inter-rater reliability session. Consent was obtained by telephone and phase 3 interviews conducted in the home of the informants throughout England (DISCO, ADI-R).

Case vignette evaluation of second-phase sample subgroup

Four hundred case vignettes were prepared for a second calibration exercise. Each vignette included a full record of the ADOS-4 (where available), together with information from the first survey phase: the AQ-20 scores, relevant information on sociodemographics, social functioning, adverse life experiences, scores on the SCID-II, Adult ADHD Screen (ASRS) and the Clinical Interview Schedule Revised (CIS-R; McManus *et al*. 2009). Six clinicians were each supplied with 100 case vignettes in batches of 50 each. Each vignette was rated independently and then rated again following discussion to make a final consensus rating. Each clinician was asked to rate the probability of the respondent having an ASD on a scale: absent, possible, probable, definite (ASD).

Data analysis

The survey data were weighted to take account of non-response, so that the results were representative of the household population aged $\geqslant$16 years (Brugha *et al*. 2009; McManus *et al*. 2009). Phase 2 weights were designed to generate condition-specific datasets that are representative of the population 'eligible' on that particular condition.

Study 1 tested the sensitivity and specificity of the ADOS-4 (phase 2), identifying the most clinically appropriate and least biased caseness threshold using two approaches. Study 1*a* used standardized ASD assessments (phase 3) of development and current functioning and diagnostic algorithms based on the DISCO (Wing & Gould, 1979) and ADI-R (Lord *et al*. 1994). Consensus clinical case vignettes made use of phase 1 and 2 survey data to also inform the calibration of ADOS-4 (Study 1*b*). (Study 2 tested the AQ-20 self-report questionnaire as a statistical predictor of a survey diagnosis of adult ASD based on ADOS-4.)

We performed AUC analyses and tests of sensitivity and specificity in Study 1*a* to assess (i.e. to 'calibrate') the agreement between the ADOS-4 threshold and the DISCO diagnosis (for which the AUC analyses are central) and to estimate prediction of autism as a binary outcome. The level of agreement was based on κ (Cohen, 1960). To compare the ADOS-4 and DISCO, a range of thresholds from 5+ to 13+ were applied to the ADOS-4 total score including that recommended for non-specific PDD (ADOS 7+) and for ASD (ADOS 10+). We also estimated agreement between the DISCO and the ADI-R.

Case vignette data were split into six parts of approximately 50 vignettes each for Study 1*b*. Each assessor worked with one other assessor and assessed 50 vignettes in a first batch. In a second batch of 50 vignettes, each assessor took a different partner. Because the second batch ranking could be an 'improvement' on the first due to practice effects, both batches were compared to evaluate the null hypothesis of no difference between the first and second

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

Table 1. *Age and sex profile: full survey sample (AQ-20 completed), phase 2 (ADOS completed)*

Age group (years)	Phase 1 (AQ-20 only) ($n=7353$)		Phase 2 (ADOS-4 completed) ($n=618$)	
	M (%)	F (%)	M (%)	F (%)
16–24	48	52	45	55
25–34	40	60	40	60
35–44	43	57	50	50
45–54	44	56	49	51
55–64	45	55	51	49
65–74	45	55	57	43
⩾75	39	61	52	48

AQ-20, 20-item Autism-Spectrum Quotient; ADOS, Autism Diagnostic Observation Schedule; M, male; F, female.

ranked batches. Each assessor ranked each vignette on a four-point scale ranging from 'no' through 'possible' and 'probable' to 'definite'. The two assessors within each pair then discussed and, if possible, agreed on a consensus rating for each vignette. The response categories were merged by combining 'no' and 'possible' into one category, and 'probable' and 'definite' into another category, yielding a binary variable based on diagnosis probability. The agreement between each of the six pairs of raters was based on unweighted κ. Agreement between batches of ratings (the first set of three and the second set of three) was assessed using the χ^2 test for independence. A further check was carried out to see whether the extent of observed agreement was affected by the presence/absence of ADOS-4 information, also using the χ^2 test for independence. Agreement between consensus ratings for the case vignettes and the ADOS 10+ and also the ADOS 7+ thresholds was based on unweighted κ.

In Study 2 we calculated the sensitivity and specificity of the AQ-20 as a predictor of the ADOS 10+ threshold recommended by the ADOS-4 developers (Lord *et al*. 2002). We estimated the best combination of sensitivity and specificity and their sum total. These estimates took into account the complex sample design used for the survey.

Results

As reported previously (Brugha *et al*. 2011), at the phase 1 interview, 57% of those eligible agreed to take part in an interview. Phase 2 interviews were conducted with 630 of those selected from phase 1 (74%); 618 ADOS interviews were completed (Fig. 1). There was no difference in the age and sex profiles of the two samples interviewed in the first two phases of the survey (Table 1).

Table 2. *Relationship of DISCO and ADI-R informants to survey respondents*[a]

Relationship	*n* (%)
Parent	25 (45)
Partner	14 (25)
Sibling	6 (11)
Other relative	7 (12)
Friend	3 (5)
Ex-partner	1 (2)

DISCO, Diagnostic Interview Schedule for Social and Communication Disorders; ADI-R, Autism Diagnostic Interview – Revised.

[a] Includes all phase 3 respondents who completed DISCO and ADI-R assessments ($n=56$).

Table 3. *DISCO and ADI-R informants by sex and age*[a]

	Completers *n* (%)	Non-completers *n*
Gender		
Males	41 (65)	24
Females	15 (43)	22
Age (years)		
16–34	12 (52)	12
35–54	18 (50)	18
55–74	18 (64)	10
⩾75	8 (73)	3

DISCO, Diagnostic Interview Schedule for Social and Communication Disorders; ADI-R, Autism Diagnostic Interview – Revised.

[a] Includes all phase 3 respondents for whom complete covariate data are available ($n=56$).

Significance of gender difference: $\chi^2=4.54$, $p=0.03$.

Significance of difference according to age: $\chi^2=2.7$, $p=0.45$.

Fifty-six completed DISCO and ADI-R phase 3 assessments were achieved (Table 2, Fig. 1). The phase 3 outcomes are set out in Tables 2 and 3. It proved easier to find and interview informants of male than of female respondents (Table 3). Twenty-four (44%) informants knew of the respondent's early development. Twenty-seven (77%) informants had at least weekly face-to-face contact with the respondent. Eighty-four per cent of informants were immediate family members and almost half were parents. There was no significant effect of age on completion of these assessments.

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

Table 4. *Comparison of DISCO and ADOS-4 (Study 1a)*[a]

ADOS cut-point	AUC	Sensitivity	Specificity	κ	DISCO positive, ADOS positive (TP)	DISCO negative, ADOS negative (TN)	DISCO positive, ADOS negative (FN)	DISCO negative, ADOS positive (FP)
5+	0.78 (0.70–0.85)	1.0 (0.59–1.0)	0.55 (0.40–0.7)	0.23 (0.07–0.4)	7	27	0	22
6+	0.78 (0.70–0.85)	1.0 (0.59–1.0)	0.55 (0.40–0.7)	0.23 (0.07–0.4)	7	27	0	22
7+	0.87 (0.80–0.93)	1.0 (0.59–1.0)	0.73 (0.58–0.85)	0.41 (0.18–0.64)	7	36	0	13
8+	0.90 (0.84–0.96)	1.0 (0.59–1.0)	0.80 (0.65–0.9)	0.49 (0.24–0.74)	7	39	0	10
9+	0.92 (0.86–0.97)	1.0 (0.59–1.0)	0.84 (0.7–0.93)	0.56 (0.3–0.82)	7	41	0	8
10+	0.93 (0.87–0.98)	1.0 (0.59–1.0)	0.86 (0.72–0.94)	0.60 (0.34–0.86)	7	42	0	7
11+	0.94 (0.89–0.99)	1.0 (0.59–1.0)	0.88 (0.75–0.96)	0.64 (0.38–0.9)	7	43	0	6
12+	0.82 (0.63–1.0)	0.71 (0.37–1.0)	0.92 (0.80–0.98)	0.56 (0.25–0.88)	5	45	2	4
13+	0.76 (0.55–0.96)	0.57 (0.18–0.91)	0.94 (0.83–0.99)	0.51 (0.16–0.86)	4	46	3	3

DISCO, Diagnostic Interview Schedule for Social and Communication Disorders; ADOS-4, Autism Diagnostic Observation Schedule, module-1; AUC, area under the curve; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives.

[a] Includes all phase 3 respondents for whom complete covariate data are available ($n=56$).

95% confidence intervals are in parentheses.

Validity of ADOS-4 in the community (*Study 1*)

Comparison of ADOS and DISCO (*Study 1a*)

In the field, less than half of the informants could provide early developmental information (Table 2), which limited the use of the ADI-R even more than the DISCO because the former requires information on development at age 4–5 years specifically. κ for agreement between the ADI-R and DISCO on ASD case identification was 0.70 (95% CI 0.39–1.0). DISCO assessments carried out with key informants yielded cases with ASD confirmed on the basis of current behaviour and, in approximately half the interviews, based also on childhood development. Fifty-six people had both an ADOS score and a DISCO assessment, together with complete information on age and gender (Table 4). Seven DISCO assessments were positive for ASD (Wing *et al*. 2002) and 45 were negative. Based on the SCAN data, no respondents positive for ASD (ADOS 10+) had been a case of psychosis.

The agreement between ADOS and DISCO assessments of each person was computed using unweighted κ statistics (Table 4). This was carried out for ADOS 7+ and also ADOS 10+ (and for all thresholds from 5+ to 13+). Seven people were assessed as ASD positive by the DISCO, 12 by the ADOS 10+ criterion, and 16 by the ADOS 7+ criterion (Table 4).

According to Table 4, for the purposes of calibration (assessing the agreement between the ADOS-4 and the DISCO diagnosis), the optimal ADOS-4 threshold, based on AUC analyses, was 10+ or 11+ ($\kappa=0.64$ for 11+). The ADOS-4 total score threshold at which the number of false positives and false negatives was closest to equal, that is where the DISCO and the ADOS-4 assessments were unbiased, was $\geqslant 13$ ($\kappa=0.51$). κ was maximum at the 11+ ADOS threshold. Identical analyses using the ADI-R instead of the DISCO produced almost the same findings with respect to the optimal ADOS-4 threshold.

Relationship of ADOS results to case vignette ratings (*Study 1b*)

A weighted κ, to assess agreement between case vignette raters, gave values ranging from 0.19 to 0.88 (Appendix). Comparison of dichotomized ratings (as in 'no' and 'possible, probable, definite') gave values from 0.19 to 0.88. As one clinician had not awarded a higher rating than 'possible', only five pairs could be compared for the categories 'no/possible versus' probable/definite' giving κ values ranging from 0.38

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

Table 5. *Comparison of vignette consensus ratings and ADOS (Study 1b)*[a,b]

ADOS cut-point	AUC	Sensitivity	Specificity	κ	Vignette positive, ADOS positive (TP)	Vignette negative, ADOS negative (TN)	Vignette positive, ADOS negative (FN)	Vignette negative, ADOS positive (FP)
5+	0.884 (0.853–0.914)	1.0 (0.69–1.0)	0.77 (0.70–0.83)	0.25 (0.12–0.38)	10	145	0	44
6+	0.844 (0.741–0.946)	0.90 (0.55–1.0)	0.79 (0.72–0.85)	0.24 (0.1–0.38)	9	149	1	40
7+	0.837 (0.703–0.970)	0.80 (0.44–0.98)	0.87 (0.81–0.92)	0.33 (0.14–0.52)	8	165	2	24
8+	0.852 (0.720–0.985)	0.80 (0.44–0.98)	0.90 (0.85–0.95)	0.40 (0.19–0.61)	8	171	2	18
9+	0.816 (0.664–0.967)	0.70 (0.34–0.94)	0.93 (0.88–0.97)	0.43 (0.2–0.66)	7	176	3	13
10+	0.818 (0.667–0.969)	0.70 (0.34–0.94)	0.94 (0.89–0.97)	0.45 (0.21–0.68)	7	177	3	12
11+	0.774 (0.612–0.935)	0.60 (0.26–0.88)	0.95 (0.90–0.98)	0.43 (0.18–0.67)	6	179	4	10
12+	0.679 (0.518–0.840)	0.40 (0.12–0.74)	0.96 (0.91–0.99)	0.33 (0.06–0.59)	4	181	6	8
13+	0.684 (0.523–0.845)	0.40 (0.12–0.74)	0.97 (0.93–0.99)	0.37 (0.09–0.65)	4	183	6	6

ADOS, Autism Diagnostic Observation Schedule; AUC, area under the curve; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives.

[a] Based only on ADOS present vignette ratings. Consensus ratings split at 0,1 *v.* 2,3.

[b] Includes all phase 3 respondents for whom complete covariate data are available ($n=199$).

95% confidence intervals are in parentheses.

to 0.79. The weighted κ method produced the best agreement of all the methods used.

There was no significant difference in the proportion of agreements between pairs of raters whether or not there was an ADOS-4 report and score present in the vignette information pack: 82% compared with 84% ($\chi^2=0.33$, $p=0.56$).

We used the clinician consensus ratings to examine their relationship with the ADOS-4; 199 subjects had both an ADOS score and a vignette assessment (Table 5). Ten of these 199 people were assessed as ASD positive by the vignette raters, 19 using the ADOS-4 10+ criterion and 32 by the ADOS-4 7+ criterion (scoring $\geqslant 7$ on the ADOS scale). There was significant and fair agreement between ADOS 7+ and vignette rating ($\kappa=0.33$, $p<0.0001$) and moderate significant agreement between ADOS 10+ and vignette rating ($\kappa=0.45$, $p<0.0001$) (Table 5).

According to Table 5, for the purposes of calibration against the consensus ratings, the optimal ADOS-4 threshold, based on AUC analyses and κ combined, is $\geqslant 8$ ($\kappa=0.4$, AUC = 0.85). The ADOS total score threshold at which the number of false positives is closest to the number of false negatives, that is where the clinical vignettes and the ADOS assessments are unbiased, is $\geqslant 12$ or $\geqslant 13$ ($\kappa=0.37$). κ achieves its maximum at the 10+ ADOS threshold (0.45). The ADOS cut-points were also ranked for κ and AUC and the ranks for each cut-point were summed. The best performing cut-points were the ones with the lowest summed ranks, which were 8+ and also 10+.

Considering the DISCO and case vignette analyses together, it was concluded that the 7+ threshold on the ADOS can be recommended for subthreshold community cases and 10+ can be recommended for definite community cases of ASD.

Validation of AQ-20 in the community (*Study 2: relationship of AQ-20 to ADOS-4 findings*)

At phase 2, 617 people had complete AQ-20 and ADOS-4 data. A positive correlation of 0.24 ($p<0.0001$) was found between the continuous AQ-20 (total) score and the continuous ADOS-4 total score. As a predictor of the ADOS 10+ threshold, the best combination of sensitivity and specificity was 0.73 and 0.62 respectively. The sum of these is 1.36: McNamee (2002) recommends a minimum sum of 1.60 if a combination of

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

a phase 1 screening tool and a phase 2 clinical measure is to be considered a cost-efficient approach.

Discussion

The present validation study is based on the first large-scale population survey of ASD in adults (Brugha *et al*. 2009). Our findings broadly support using the ADOS-4 developers' recommended diagnostic cut-points of ADOS-4 total score 7+ (non-specific ASD) and 10+ (autism) verified in clinic patients, which we would term respectively as subthreshold adult community cases (7+ threshold) and definite ASD adult cases in the community (10+ threshold). However, the AQ-20 score was found to have a low correlation with the ADOS-4 score, with poor sensitivity and specificity for the ADOS-4 binary outcome, and therefore is of limited use as a screening tool in general population surveys.

The ADI-R and the adult version of the ADOS used here approximate closely to the childhood diagnostic instruments used in the large-scale surveys of children by Baird *et al*. (2006) and Baron-Cohen *et al*. (2009). The assessment methods used in our two studies are therefore of the highest standard achievable at this point in time.

Various issues can be considered in determining a threshold that is optimal in relation to a reference assessment, hence our use of the term calibration. There are statistical issues and there are clinical issues based on judgements of the use of instruments in day-to-day clinical practice. Where instrument thresholds are shown to differ, this could be termed statistically as bias because, in determining prevalence, what is wanted, ideally, is the same proportion being deemed cases on both instruments. The 13+ ADOS-4 threshold provided optimally low bias on both referencing comparisons but sensitivity was particularly poor on the vignette analyses and both κ and AUC were poorer.

The moderate agreement between clinical vignette raters and the ADOS-4 findings suggests that, albeit with limited information, such specialist practitioners use a wide range of diagnostic thresholds. Thus, the vignette results (Table 5) are less clear-cut than the DISCO results (Table 4). The vignettes support a range of ADOS-4 thresholds from 7+ to 10+. (The DISCO results table pointed to a higher 11+ ADOS-4 threshold as optimal based on AUC analysis and κ.) Our clinicians may have placed different amounts of weight on each source of information used. It could be argued that clinicians have more complex and difficult judgements to make compared with trained research interviewers; clinicians have to weigh up many different factors including, possibly, whether a case might benefit from having a diagnosis and thus access to support and care, which would never be a consideration for research interviewers.

The statistical analyses possibly indicate a higher threshold than 10+. However, differences between each threshold are marginal. A range of thresholds from 7+ to 10+ was favoured by the clinicians, who also had access to textual descriptions of current behaviours observed by the ADOS-4 assessors. So-called 'false negatives' began to be seen from as low as ADOS 6+ with reference to the vignette assessments and 13+ with reference to the DISCO. DISCO ratings may have been attenuated by the high age of respondents and parental informants so that fewer assessments fulfilled diagnostic criteria. These together would argue against imposing a threshold any higher than 10+.

It is also important to take into consideration the thresholds previously recommended by the developers of the ADOS-4, namely 7+ (autism spectrum cut-off) and 10+ (autism cut-off). There would need to be a strong case and clear evidence to lead to an increase in this threshold, which was developed in the context of clinical practice and with reference to recommendations based on DSM-IV diagnostic criteria. It was therefore concluded that the 7+ threshold on the ADOS should be used for subthreshold community cases and 10+ for definite community cases of ASD. The 10+ threshold is recommended to be used therefore in estimating prevalence rates in adults in a community survey. However, we would also recommend that investigators report their findings at a range of thresholds, as we have (Brugha *et al*. 2011).

Several study limitations require discussion. There are no objective tests for diagnosing autism; recognition and diagnosis depends on information about the pattern of behaviour and skills observed. Clinical referencing of the ADOS-4 was carried out within the limitations set by the survey fieldwork context and the willingness of adults, including elderly participants, to help with the study. There was a relatively poor rate of respondent cooperation at the survey first and third phases although more cooperation at phase 2, but in separately reported analyses we find no evidence of significant non-response bias in probability of ASD (Brugha *et al*. 2011). Clinicians were shown the available ADOS subscale scores and interviewer text descriptions of observed behaviour as part of the consensus case vignette task and were therefore not blind to the reference instrument. This did not prevent them making diagnostic judgements that were more discordant with the ADOS-4 findings compared to comparisons between the DISCO and ADOS-4. Nevertheless, such 'best estimate clinical diagnosis' procedures, although less transparent, are valued in clinical practice (Lord *et al*. 2002). We also kept

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

separate the case vignette analyses because such procedures cannot be described sufficiently precisely to facilitate independent replication.

The informant-based DISCO (and ADI-R) assessments may also have been limited by the availability of adults who knew the respondents closely and their ability to recall details of childhood development decades previously. Approximately half of the case vignettes did not include second-phase ADOS-4 findings and in practice proved unusable because clinicians had too little information. One important feature of ASD, repetitive and stereotyped behaviour, was rarely observed in adults during a 90-min direct examination and we would concur with the exclusion of ratings of such behaviours from the module 4 diagnostic algorithm for the ADOS (Lord *et al*. 2002).

In contrast to clinic settings, it proved difficult to build up a substantial sample of key informants available and willing to take part in the lengthy DISCO and ADI-R assessment within this general population survey. Therefore, the phase 3 sample provided indicative rather than clearly representative findings.

Although the AQ-20 was developed in conjunction with the originators of the full AQ, our findings can only apply to the abbreviated 20-item version. The AQ-20 self-report screening questionnaire score was found to have a low correlation (0.24; $p < 0.0001$) with the continuous ADOS-4 total score and unsatisfactory sensitivity and specificity with the ADOS 10+ threshold. It was not possible to predict confidently which of the phase 1 respondents with AQ-20 scores of $\geqslant 5$ had ASD unless they had been assessed on the ADOS-4 in phase 2. Hence the final prevalence estimates obtained were based only on data from phase 1 respondents with an AQ-20 score below 5 together with all who completed phase 2.

Our results with the AQ-20 suggest limited use as a screening tool in general population surveys and the need for further work to improve screening methods for such populations. Fortunately, in our general population survey, we had sufficient ADOS-4 data on respondents across the full range of AQ-20 scores so that by weighting our findings we were able to produce reliable prevalence estimates. Until a better screening tool has been developed and tested in the community, the AQ-20 remains the only fully evaluated self-report measure of ASD that can be used in surveys of adults. It may still have value in screening individuals who are seeking a specialist assessment (Brugha *et al*., unpublished observations), provided other forms of mental disorder (co-morbidity) are also assessed clinically.

In conclusion, methods have been developed that are feasible and that provide the possibility of generating data on the epidemiology of the ASD phenotype in adults capable of taking part in a general population survey interview. The first of two validation studies confirmed the use of the previously established 10+ threshold on the ADOS-4 in community surveys. The second study found the AQ-20 to be of limited value as a self-report assessment in general population research. As this is the first methodological evaluation of its kind in adults in the autism field, it is to be hoped that future research could build on and improve the progress achieved with the present survey evaluation.

Appendix

Agreement between vignette raters

Rater pair	κ (1,2 *v*. 3,4)	κ (1 *v*. 2,3,4)	Weighted κ
AB	N.A.	0.21	0.19
CB	0.49	0.62	0.73
CD	0.79	0.31	0.59
DE	0.38	0.58	0.74
EF	0.66	0.88	0.90
AF	0.66	0.19	0.63

N.A., Not applicable.

Rater pair refers to the six clinical vignette raters coded A to F.

Acknowledgements

Funding was provided by The National Health Service (NHS) Information Centre for Health and Social Care and the Department of Health, London, UK; The National Institute for Health Research (NIHR) and the Department of Health Policy Research Programme, London, UK.

The authors are solely responsible for the findings reported here and acknowledge suggestions and advice received during the development of this research from the following: Dr L. Wing, Dr J. Gould, Professor S. Baron-Cohen, Dr S. Wheelwright, Dr A. Wakabayashi, Professor A. Le Couteur, Prof. C. Lord, and the APMS group. (APMS survey website: www.mentalhealthsurveys.co.uk/.)

Declaration of Interest

None.

References

APA (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC.

Asperger H (1991). 'Autistic psychopathy' in childhood. In *Autism and Asperger Syndrome* (ed. U. Frith), pp. 37–92. Cambridge University Press: Cambridge.

https://doi.org/10.1017/S0033291711001292 Published online by Cambridge University Press

Baird G, Simonoff E, Pickles A, Chandler S, Loucas T, Meldrum D, Charman T (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *Lancet* **368**, 210–215.

Baron-Cohen S, Scott FJ, Allison C, Williams J, Bolton P, Matthews FE, Brayne C (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *British Journal of Psychiatry* **194**, 500–509.

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders* **31**, 5–17.

Brugha TS, Bebbington PE, Jenkins R (1999). A difference that matters: comparisons of structured and semi-structured diagnostic interviews of adults in the general population. *Psychological Medicine* **29**, 1013–1020.

Brugha TS, McManus S, Bankart J, Scott F, Purdon S, Smith J, Bebbington P, Jenkins R, Meltzer H (2011). Epidemiology of autism spectrum disorders in adults in the community in England. *Archives of General Psychiatry* **68**, 459–466.

Brugha TS, McManus S, Meltzer H, Smith J, Scott F, Purdon S, Harris J, Meltzer H (2009). *Autism Spectrum Disorders in adults living in households throughout England – report from the Adult Psychiatric Morbidity Survey 2007*. The NHS Information Centre: Leeds.

Brugha TS, Meltzer H (2008). Measurement of psychiatric and psychological disorders and outcomes in populations. In *International Encyclopedia of Public Health* (ed. K. Heggenhougen and S. Quah), pp. 261–272. Academic Press: San Diego.

Brugha TS, Nienhuis FJ (1998). *SCAN-SF. A Survey Form of the Present State Examination and SCAN: Supplementary schedule for lay interviewers*. World Health Organization: Geneva.

Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry* **166**, 540–556.

Cohen J (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement* **20**, 37–46.

Elashoff J (1997). *nQuery Advisor® Release 2.0*. Statistical Solutions: Boston, MA.

Fombonne E (2003). Epidemiological surveys of autism and other pervasive developmental disorders: an update. *Journal of Autism and Developmental Disorders* **33**, 365–382.

Fombonne E (2005). Epidemiology of autistic disorder and other pervasive developmental disorders. *Journal of Clinical Psychiatry* **66** (Suppl. 10), 3–8.

Frith U (1991). *Autism and Asperger Syndrome*. Cambridge University Press: Cambridge.

Haro JM, Arbabzadeh-Bouchez S, Brugha TS, de Girolamo G, Guyer ME, Jin R, Lepine JP, Mazzi F, Reneses B, Vilagut G, Sampson NA, Kessler RC (2006). Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *International Journal of Methods in Psychiatric Research* **15**, 167–180.

Howlin P, Goode S, Hutton J, Rutter M (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry* **45**, 212–229.

Jenkins R, Meltzer H, Bebbington P, Brugha T, Farrell M, McManus S, Singleton N (2009). The British Mental Health Survey Programme: achievements and latest findings. *Social Psychiatry and Psychiatric Epidemiology* **44**, 899–904.

Kanner L (1943). Autistic disturbance of affective contact. *Nervous Child* **2**, 217–250.

Knapp M, Romeo R, Beecham J (2007). *Economic Consequences of Autism in the UK*. Mental Health Foundation and Autism Speaks: London.

Lord C, Rutter M, DiLavore PC, Risi S (2002). *Autism Diagnostic Observation Schedule. ADOS Manual*. Western Psychological Services: Los Angeles.

Lord C, Rutter M, Le Couteur A (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders* **24**, 659–685.

McManus S, Meltzer H, Brugha T, Bebbington P, Jenkins R (2009). *Adult Psychiatric Morbidity in England, 2007. Results of a Household Survey*. The NHS Information Centre for Health and Social Care: London.

McNamee R (2002). Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in Medicine* **21**, 3609–3625.

WHO (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization: Geneva.

Williams JB, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Pope HGJ, Rounsaville B (1992). The Structured Clinical Interview for DSM-III-R (SCID). II. Multisite test-retest reliability. *Archives of General Psychiatry* **49**, 630–636.

Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablensky A, Regier D, Sartorius N (1990). SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* **47**, 589–593.

Wing L (1997). The autistic spectrum. *Lancet* **350**, 1761–1766.

Wing L, Gould J (1979). Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification. *Journal of Autism and Developmental Disorders* **9**, 11–29.

Wing L, Leekam SR, Libby SJ, Gould J, Larcombe M (2002). The Diagnostic Interview for Social and Communication Disorders: background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry* **43**, 307–325.